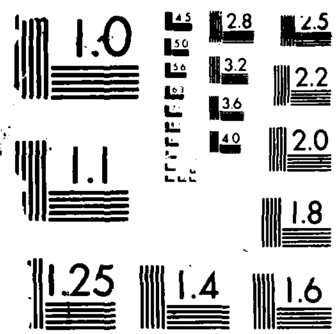
AD-A186 584   STRONG CONSISTENCY OF CERTAIN INFORMATION THEORETIC CRITERIA FOR MODEL SE..(U) PITTSBURGH UNIV PA CENTER FOR MULTIVARIATE ANALYSIS R NISHII ET AL. DEC 86   1/1

UNCLASSIFIED   TR-86-42 AFOSR-TR-87-1005 F49620-85-C-0008   F/G 12/3   NL



MICROCOPY RESOLUTION TEST CHART

AFOSR-TR- 87-1005

Center for Multivariate Analysis

University of Pittsburgh

87 9 24 138

STRONG CONSISTENCY OF CERTAIN INFORMATION THEORETIC CRITERIA FOR MODEL SELECTION IN *CALIBRATION, DISCRIMINANT ANALYSIS AND* CANONICAL CORRELATION ANALYSIS*

R. Nishii
University of Pittsburgh and Hiroshima University

Z.D. Bai and P.R. Krishnaiah
University of Pittsburgh

December 1986

Technical Report No. 86-42

Center for Multivariate Analysis
Fifth Floor Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

* Research supported by the Air Force Office of Scientific Research (AFOSC) under Contract F49620-85-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

STRONG CONSISTENCY OF CERTAIN INFORMATION THEORETIC CRITERIA FOR MODEL SELECTION IN CALIBRATION, DISCRIMINANT ANALYSIS AND CANONICAL CORRELATION ANALYSIS*

R. Nishii, Z.D. Bai and P.R. Krishnaiah

ABSTRACT

~~In~~ this paper, ~~the~~ authors show that the criteria for model selection based upon efficient detection (ED) criterion are consistent for certain problems in multivariate calibration, discriminant analysis and canonical correlation analysis. These results will be proved under mild conditions on the underlying distribution.

Keywords and phrases: AIC, information criteria, law of iterated logarithm, multivariate analysis.

AMS 1980 Subject Classification: Primary 62J05; Secondary 62H20, 62H30

* Research supported by the Air Force Office of Scientific Research (AFOSC) under Contract F49620-85-C-0008.

AD-A186 584

(1)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

DTIC FILE COPY

1. REPORT NUMBER AFOSR-TR- 87-1005	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Strong consistency of certain information theoretic criteria for model selection in calibration, discriminant analysis and canonical correlation analysis		5. TYPE OF REPORT & PERIOD COVERED Journal ~~Technical~~ - December 1986
		6. PERFORMING ORG. REPORT NUMBER 86-42
7. AUTHOR(s) R. Nishii, Z.D. Bai and P.R. Krishnaiah		8. CONTRACT OR GRANT NUMBER(s) F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Multivariate Analysis, Fifth Floor Thackeray Hall, University of Pittsburgh, Pittsburgh, PA 15260		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304 A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research, Department of the Air Force, Bolling Air Force Base, DC 20332 NM		12. REPORT DATE December 1986
		13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Same as 11		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DTIC ELECTE OCT 16 1987 S D

18. SUPPLEMENTARY NOTES

CD

19 *Keywords and phrases:* AIC, information criteria, law of iterated logarithm, multivariate analysis.

20 ABSTRACT

In this paper, the authors show that the criteria for model selection based upon efficient detection (ED) criterion are consistent for certain problems in multivariate calibration, discriminant analysis and canonical correlation analysis. These results will be proved under mild conditions on the underlying distribution.

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

1. INTRODUCTION

In the area of model selection, various procedures have been proposed in the literature and their properties are examined. In this paper we consider a generalized information criterion (GIC) obtained by the information theoretic approach. According to this procedure, we find the model which minimizes

$$\text{GIC} = -2 \log L(\hat{\theta}) + c_N p$$

where $L(\hat{\theta})$ is the maximized likelihood and p is the number of parameters. Akaike (1973) proposed to take $c_N \equiv 2$, and Rissanen (1978) and Schwartz (1978) proposed $c_N = \log N$ where N denotes the sample size (see also Akaike (1978) and Hannan and Quinn (1979)). Recently Zhao, Krishnaiah and Bai (1986) considered the GIC such that (i) $\lim_{N\to\infty} c_N/N = 0$ and (ii) $\lim_{N\to\infty} c_N/\log\log N = +\infty$. The above criterion is sometimes referred to as efficient detection (ED) criterion. They used the criterion for the determination of the number of signals under a signal processing model.

In the present paper, we propose to use the ED criterion for certain problems of multivariate analysis. Sometimes statistician is expected to predict the explanatory variables using some of the response variables under the multivariate regression model. This problem is treated in Section 2 by using the ED criterion, and its consistency is established. Here we may note that Nishii (1986) pointed out the inconsistency of Akaike's AIC in calibration. In Section 3 we discuss the selection of variables in discriminant analysis. Our interest is to find the variables which contribute for discrimination between the populations. Section 4 is concerned with the selection of variables in canonical correlation analysis, i.e., among two sets of variables we want to find which subsets are important for studying the association between two sets. The investigations for the above cases are made under a mild condition on the underling distribution.

2. MULTIVARIATE CALIBRATION

Let q explanatory variables $\underset{\sim}{x} \equiv (x_1, \ldots, x_q)'$ and p response variables $\underset{\sim}{y} \equiv (y_1, \ldots, y_p)'$ have the linear relation:

$$\underset{\sim}{y} = \underset{\sim}{\alpha} + \beta'\underset{\sim}{x} + \underset{\sim}{e} \tag{2.1}$$

where $\underset{\sim}{e}$ follows $N_p[\underset{\sim}{0},\Sigma]$, $\underset{\sim}{\alpha}$: $p \times 1$, β: $q \times p$ and Σ: $p \times p$ are parameters. Suppose we are interested in estimating $\underset{\sim}{x}$ by using observed $\underset{\sim}{y}$. If all parameters are known, the maximum likelihood estimate of the unknown explanatory variables $\underset{\sim}{x}$ is obtained by

$$\hat{\underset{\sim}{x}} = (\beta\Sigma^{-1}\beta')^{-}\beta\Sigma^{-1}(\underset{\sim}{y} - \underset{\sim}{\alpha}), \tag{2.2}$$

where $(\beta\Sigma^{-1}\beta')^{-}$ is a G-inverse of $\beta\Sigma^{-1}\beta'$. However, if the last column of $\beta\Sigma^{-1}$ is zero vector, the response variable y_p would supply no additional information on $\underset{\sim}{x}$ in the multivariate linear model (see §4 of Rao (1973)). Hence, we want to obtain the best subset of response variables such that each of its elements has some information. For this problem, criteria based on information theory can be used. For a review of the literature on multivariate calibration, the reader is referred to Brown (1982).

Let J be a subset of indices of response variables $\{1, \ldots, p\}$. We say that "the assumed model is J" when we regard that y_j $(j \in J)$ provides information for $\underset{\sim}{x}$ whereas $y_{j'}$ $(j' \notin J)$ does not. We assume the existence of the true model $\{1, \ldots, p_t\} = J_t$ but it is unknown and let $p_t \leq p$. This assumption is equivalent to

$$\beta_J\Sigma_{JJ}^{-1}\beta_J' \begin{cases} = \beta_t\Sigma_{tt}^{-1}\beta_t' & \text{if} \quad J \supseteq J_t \\ \leqq \beta_t\Sigma_{tt}^{-1}\beta_t' & \text{if} \quad J \not\supseteq J_t \end{cases} \tag{2.3}$$

and $\operatorname{tr}\beta_J\Sigma_{JJ}^{-1}\beta_J' < \operatorname{tr}\beta_t\Sigma_{tt}^{-1}\beta_t'$ if $J \not\supseteq J_t$ where β_J: $q \times \#J$ and Σ_{JJ}: $\#J \times \#J$ are submatrices of β: $q \times p$ and Σ: $p \times p$ corresponding to a subset J, #J denotes the number of elements of J, and β_t: $q \times p_t$ and Σ_{tt}: $p_t \times p_t$ are, corresponding to J_t, are similarly defined (see McKay (1977) and Fujikoshi (1983)).

When all parameters are unknown, and N independent observations $\underset{\sim}{y}_i$ at $\underset{\sim}{x}_i$ (i = 1, ..., N) with the relationship (2.1) are given, we use the estimates of $\underset{\sim}{\alpha}$, β and $N\Sigma$ as

$$\underset{\sim}{a} = \bar{\underset{\sim}{y}} - B'\bar{\underset{\sim}{x}}, \quad B = S_{XX}^{-1}S_{XY} \quad \text{and} \quad S = S_{YY} - B'S_{XX}B \tag{2.4}$$

where

$$\begin{pmatrix} \bar{\underset{\sim}{x}} \\ \bar{\underset{\sim}{y}} \end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \underset{\sim}{x} \\ \underset{\sim}{y} \end{pmatrix}, \quad \begin{pmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{pmatrix} = \sum_{i=1}^{N}\begin{pmatrix} \underset{\sim}{x}_i - \bar{\underset{\sim}{x}} \\ \underset{\sim}{y}_i - \bar{\underset{\sim}{y}} \end{pmatrix}\begin{pmatrix} \underset{\sim}{x}_i - \bar{\underset{\sim}{x}} \\ \underset{\sim}{y}_i - \bar{\underset{\sim}{y}} \end{pmatrix}'. \tag{2.5}$$

Note that S and $B'S_{XX}B$ follow the Wishart distribution $W_p[N-q-1, \Sigma]$ and the noncentral Wishart distribution $W_p[q, \Sigma; \beta'S_{XX}\beta]$ respectively. The likelihood ratio for the model J against the full model $J_f \equiv \{1, \ldots, p\}$ for N calibration samples is expressed by Fujikoshi and Nishii (1986). Hence,

$$G_N(J) \equiv GIC(J) - GIC(J_f) = \Lambda(J) - q(p - \#J)c_N \tag{2.6}$$

where

$$\Lambda(J_f; J) = N \log \frac{|S_{JJ}|\,|S + B'S_{XX}B|}{|S|\,|S_{JJ} + B_J'S_{XX}B_J|}. \tag{2.7}$$

We select the model $\hat{J}_N$ such that

$$G_N(\hat{J}_N) = \min_J G_N(J). \tag{2.8}$$

Recall the criterion function (2.6) is derived when $\underset{\sim}{y}_i$ are normally dis-

tributed. However, we apply this procedure when we relax the assumption of normality. Nishii (1986) studied the asymptotic behavior of the AIC for the case $c_N \equiv 2$ in (2.6) under a weak assumption and he showed that the AIC is not consistent in multivariate calibration problem. If we use the ED criterion, c_N is chosen such that

$$\text{(i)} \quad \lim_{N\to\infty} (c_N/N) = 0, \qquad \text{(ii)} \quad \lim_{N\to\infty} (c_N/\log\log N) = \infty.$$

We will show that the MDL criterion is strongly consistent under the following mild conditions:

ASSUMPTION 1. The error vectors $\underset{\sim}{e}_i$ of $\underset{\sim}{y}_i$ $(i = 1, \ldots, N, \ldots)$ are independently and identically distributed (i i.d) with

$$E\underset{\sim}{e}_1 = \underset{\sim}{0}, \quad E\underset{\sim}{e}_1\underset{\sim}{e}_1' = \Sigma \quad \text{and} \quad E(\underset{\sim}{e}_1'\underset{\sim}{e}_1)^{\gamma/2} < \infty \tag{2.9}$$

for some $\gamma \in [2, 3]$.

ASSUMPTION 2. The sequence of the vectors of explanatory variables $\{\underset{\sim}{x}_i = (x_{i1}, \ldots, x_{iq})' \mid i = 1, \ldots, N, \ldots\}$ satisfies

$$\text{(i)} \qquad 0 < mI_q \le N^{-1}S_{XX} = N^{-1}\sum_{i=1}^{N}(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}_N)(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}_N)' \le MI_q, \tag{2.10}$$

$$\text{(ii)} \qquad \sum_{i=1}^{N}|x_{ik} - \bar{x}_{Nk}|^{\gamma} \le \begin{cases} \Gamma N^{\gamma/2}(\log\log N)^{3/2}, & (2 \le \gamma < 3) \\ \Gamma N^{3/2}/\log N, & (\gamma = 3) \end{cases} \tag{2.11}$$

where $\underset{\sim}{\bar{x}}_N = N^{-1}(\underset{\sim}{x}_1 + \ldots + \underset{\sim}{x}_N) = (\bar{x}_{N1}, \ldots, \bar{x}_{Nq})'$, m, M and Γ are positive constants, and γ is given in Assumption 1. Here k runs through 1 to q.

The proof of the following lemma is given in the Appendix.

LEMMA 2.1. Under Assumptions 1 and 2, it holds that

$$T_N = \sum_{i=1}^{N} (\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}_N)\underset{\sim}{e}_i' : q \times p = O((N \log\log N)^{1/2}), \quad \text{a.s.} \tag{2.12}$$

THEOREM 2.1. Under Assumptions 1 and 2, the model selection procedure based on the ED criterion is strongly consistent in multivariate calibration problem, i.e., $\lim_{N\to\infty} \hat{J}_N = J_t$, a.s.

Proof. From Assumption 2, $S_{XX} = O(N)$. Using Lemma 2.1 and the law of iterated logarithm, we have

$$\begin{aligned} N^{-1}B'S_{XX}B &= N^{-1}\beta'S_{XX}\beta + T_N'\beta + \beta'T_N + T_N'S_{XX}^{-1}T_N \\ &= N^{-1}\beta'S_{XX}\beta + O(\ell_N), \quad \text{a.s.,} \end{aligned} \tag{2.13}$$

$$\begin{aligned} N^{-1}S &= N^{-1}(S_{YY} - B'S_{XX}B) \\ &= N^{-1}\sum_{i=1}^{N}(\underset{\sim}{e}_i - \underset{\sim}{\bar{e}}_N)(\underset{\sim}{e}_i - \underset{\sim}{\bar{e}}_N)' - N^{-1}T_N'S_{XX}^{-1}T_N \\ &= \Sigma + O(\ell_N), \quad \text{a.s.,} \end{aligned} \tag{2.14}$$

where $T_N: q \times p$ is defined in (2.12) and $\ell_N = (N^{-1}\log\log N)^{1/2}$. If $J \not\supseteq J_t$, by (2.5), (2.13) and (2.14), we have

$$G_N(J) = \mathrm{tr}\{(\beta\Sigma^{-1}\beta' - \beta_J\Sigma_{JJ}^{-1}\beta_J')S_{XX}\} - q(p - \#J)c_N + O(N^{1/2}\ell_N), \quad \text{a.s.} \tag{2.15}$$

The first term of the right hand side of (2.15) is positive by (2.3) and it increases with the order N by (2.10), which together with $\lim_{N\to\infty} N^{-1}c_N = 0$ implies

$$G_N(J) > 0 \text{ for large } N, \quad \text{a.s.} \tag{2.16}$$

On the other hand $G_N(J_f) \equiv 0$ for any N by the definition of G_N. This yields that MDL criterion asymptotically prefers J_f to J if $J \not\supset J_t$. When $J_f = J_t$, the proof follows. If $J_f \neq J_t$, at first we consider the case $J = J_f \supsetneq J_t$. Denote $S = \begin{pmatrix} S_{tt} & S_{t1} \\ S_{1t} & S_{11} \end{pmatrix}: p \times p$, $S_{tt}: p_t \times p_t$, $B = [B_t, B_1]: q \times p$, $B_t: q \times p_t$. Let $S_{11\cdot t} = S_{11} - S_{1t}S_{tt}^{-1}S_{t1}$ and define $(S + B'S_{XX}B)_{11\cdot t}$ in a similar way. Put $U = S_{XX}^{1/2}B = [U_t, U_1]: q \times p$ and $U_t: q \times p_t$. From Fujikoshi (1983), we know that

$$
\begin{aligned}
(S + B'S_{XX}B)_{11\cdot t} - S_{11\cdot t} &= (S + U'U)_{11\cdot t} - S_{11\cdot t} \\
&= (U_1 - U_tS_{tt}^{-1}S_{t1})'(I_q + U_tS_{tt}^{-1}U_t')^{-1}(U_1 - U_tS_{tt}^{-1}S_{t1}).
\end{aligned}
$$

By the law of iterated logarithm and Lemma 2.1, we have

$$
N^{-1}S_{11\cdot t} = \Sigma_{11\cdot t} + O(\ell_N), \quad \text{a.s.},
$$

$$
\begin{aligned}
U_tS_{tt}^{-1}U_t' &= N^{-1}S_{XX}^{1/2}\beta_t\Sigma_{tt}^{-1}\beta_t'S_{XX}^{1/2} + O(\ell_N), \quad \text{a.s.}, \\
&= O(1), \quad \text{a.s.}
\end{aligned}
$$

$$
\begin{aligned}
U_1 - U_tS_{tt}^{-1}S_{t1} &= S_{XX}^{1/2}\beta_1 - S_{XX}^{1/2}\beta_t\Sigma_{tt}^{-1}\Sigma_{t1} + O(N^{1/2}\ell_N), \quad \text{a.s.}, \\
&= O(N^{1/2}\ell_N), \quad \text{a.s.}
\end{aligned}
$$

The last equality follows from the relation $\beta_1 = \beta_t\Sigma_{tt}^{-1}\Sigma_{t1}$ which is obtained by (2.3). Hence

$$
\begin{aligned}
G_N(J_t) &= \Lambda(J_f, J_t) - q(p - p_t)c_N \\
&= N \log \frac{|(S + U'U)_{11\cdot t}|}{|S_{11\cdot t}|} - q(p - p_t)c_N \\
&= N \log|I_{p-p_t} + S_{11\cdot t}^{-1}\{(S + U'U)_{11\cdot t} - S_{11\cdot t}^{-1}\}| - q(p - p_t)c_N \\
&= O(\log\log N) - q(p - p_t)c_N \to -\infty, \quad (N \to \infty), \quad \text{a.s.} \qquad (2.17)
\end{aligned}
$$

because $p - p_t > 0$ and $\lim_{N\to\infty} c_N/\log\log N = +\infty$. This implies that the ED criterion will not asymptotically select the model J_f. When $J \supsetneq J_t$ following similar lines as in the above, it holds that

$$\Lambda(J_f, J) = O(\log\log N), \quad \text{a.s.}$$

Hence,

$$G_N(J_t) - G_N(J) = \Lambda(J_f, J_t) - \Lambda(J_f, J) - q(p - \#J)c_N$$

$$= O(\log\log N) - q(p - \#J)c_N \to -\infty, \quad \text{a.s.}$$

This completes the proof.

However, we must calculate $2^p - 1$ $G_N(\cdot)$'s to obtain $\hat{J}_N$ of (2.8). When p is large, this would involve extensive computation. To overcome this problem, we propose an alternate procedure, which is also based on the MDL criterion. Let $J_{-i} = \{1, \ldots, i-1, i+1, \ldots, p\}$ for $i = 1, \ldots, p$. Define

$$\tilde{J}_N = \{i \in J_f \mid G_N(J_{-i}) > 0 = G_N(J_f)\}. \tag{2.18}$$

This subset is obtained by calculating only $p+1$ $G_N(\cdot)$'s, but this is still a strongly consistent estimate of J_t. (See Zhao, Krishnaiah and Bai (1986).)

THEOREM 2.2. Under Assumptions 1 and 2, we have

$$\lim_{N\to\infty} \tilde{J}_N = J_t, \quad \text{a.s.}$$

Proof. If $i \in J_t$, then $J_{-i} \not\supseteq J_t$. By (2.15), $G_N(J_{-i})$ tends almost surely to infinity. Hence $\tilde{J}_N \ni i$ for large N, a.s. If $i \notin J_t$, then $J_{-i} \supseteq J_t$. By similar discussion as (2.17), we have

$$G_N(J_{-i}) \to -\infty \quad \text{as} \quad N \to \infty, \quad \text{a.s.}$$

This implies $i \notin \tilde{J}_N$ for large N, a.s., and this completes the proof.

3. DISCRIMINANT ANALYSIS

The discussion on multivariate calibration can be applied to the variable selection in multiple discriminant analysis. Consider $q+1$ p-variate normal populations Π_α with mean vector $\underset{\sim}{\mu}_\alpha$ and common covariance matrix Σ ($\alpha = 1, \ldots, q+1$). Assume N_α samples $\underset{\sim}{x}_{\alpha 1}, \ldots, \underset{\sim}{x}_{\alpha N_\alpha}$ are drawn from Π_α. We are interested in interpreting the differences among the $q+1$ populations in terms of only a few canonical discriminant variates.

Let Ω be the population between-groups covariance matrix as

$$\Omega = N^{-1}\sum_{\alpha=1}^{q+1} N_\alpha(\underset{\sim}{\mu}_\alpha - \bar{\underset{\sim}{\mu}})(\underset{\sim}{\mu}_\alpha - \bar{\underset{\sim}{\mu}})' : p \times p,$$

where $\bar{\underset{\sim}{\mu}} = N^{-1}\sum \underset{\sim}{\mu}_\alpha$ and $N = \sum N_\alpha$. Let J be a subset of $\{1, \ldots, p\} \equiv J_f$. We say that the model is J when unknown parameters satisfy

$$\mathrm{tr}\,\Sigma^{-1}\Omega = \mathrm{tr}\,\Sigma_{JJ}^{-1}\Omega_{JJ} > \mathrm{tr}\,\Sigma_{J'J'}^{-1}\Omega_{J'J'} \quad \text{for} \quad j' \not\supset J \tag{3.1}$$

where Ω_{JJ} and Σ_{JJ} are $\#J \times \#J$ submatrices of Ω and Σ respectively. We assume that the true model exists and denote it by $J_t = \{1, \ldots, p_t\}$. The maximum likelihood function under the model J is known (see Fujikoshi (1983)). Hence, we have

$$\begin{aligned} G_N(J) &= \mathrm{GIC}(J) - \mathrm{GIC}(J_f) \\ &= N \log \frac{|W_{JJ}|\,|W+U|}{|W|\,|W_{JJ}+U_{JJ}|} - q(p - \#J)c_N \end{aligned} \tag{3.2}$$

where

$$W = \sum_{\alpha=1}^{q+1}\sum_{i=1}^{N_\alpha}(\underset{\sim}{z}_{\alpha i} - \bar{\underset{\sim}{z}}_\alpha)(\underset{\sim}{z}_{\alpha i} - \bar{\underset{\sim}{z}}_\alpha)' : \quad , \tag{3.3}$$

$$U = \sum_{\alpha=1}^{q+1} N_\alpha (\bar{z}_\alpha - \bar{z})(\bar{z}_\alpha - \bar{z})' : p \times p \tag{3.4}$$

$\bar{z}_\alpha = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} z_{\alpha i}$, $\bar{z} = N^{-1} \sum_{\alpha=1}^{q+1} N_\alpha \bar{z}_\alpha$. Here W and U respectively denote the within group sums of squares and cross products (SP) matrices. Note that $W \sim W_p[N-q-1, \Sigma]$ and $U \sim W_p[q, \Sigma; N\Omega]$, and recall that $S \sim W_p[N-q-1, \Sigma]$ and $B'S_{XX}B \sim W_p[q, \Sigma; \beta' S_{XX}\beta]$ in (2.5). Let $\{S_{XX} = S_{XX}^{(N)}\}$ be a sequence satisfying Assumption 2 with $\gamma = 2$. Then we can find $\beta = \beta_N$; $q \times p$ such that $\beta' S_{XX} \beta = N\Omega$ since rank $\Omega \leq p, q$. Put $S = W$ and $B'S_{XX} = U$ in (2.5). This gives the correspondence between (2.5) and (3.2) except that β depends on N.

Let $\hat{J}_N$ be a subset of J_f minimizing (3.2) and let $\tilde{J}_N$ be a subset of J_f defined by (2.18) in this situation.

THEOREM 3.1. Let $z_{\alpha i} - \mu_\alpha$ ($i = 1, \ldots, N_\alpha$; $\alpha = 1, \ldots, q+1$) be i.i.d with $E(z_{\alpha i} - \mu_\alpha) = 0$ and $E(z_{\alpha i} - \mu_\alpha)(z_{\alpha i} - \mu_\alpha)' = \Sigma$. Assume that the data increases satisfying the condition

$$0 < m' < N^{-1} N_\alpha < 1 \quad (\alpha = 1, \ldots, q+1), \qquad N = \sum N_\alpha$$

where m' is a positive constant. Then both $\hat{J}_N$ and $\tilde{J}_N$ are strongly consistent estimators of J_t.

4. CANONICAL CORRELATION ANALYSIS

In this section we treat the variable selection problem in canonical correlation analysis. Let $\underset{\sim}{z} = (\underset{\sim}{x}', \underset{\sim}{y}')'$ follow $N_{p+q}[\mu, \Sigma]$ where $\underset{\sim}{x}$: $q \times 1$, $\underset{\sim}{y}$: $p \times 1$, $\underset{\sim}{\mu} = (\underset{\sim}{\mu}_X', \underset{\sim}{\mu}_Y')'$: $(p+q) \times 1$, $\underset{\sim}{\mu}_X$: $q \times 1$, $\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$: $(p+q) \times (p+q)$ and Σ_{XX}: $q \times q$. Suppose we are interested in summarizing the relationship between $\underset{\sim}{x}$ and $\underset{\sim}{y}$ by using a small number of variables. Let $I_f = \{1, \ldots, q\}$ and $J_f = \{1, \ldots, p\}$ be sets of the indices of $\underset{\sim}{x}$ and $\underset{\sim}{y}$ respectively. Consider subsets $I \subseteq I_f$ and $J \subseteq J_f$. We say that the model is (I,J) when, using submatrix Σ_{JJ} of Σ_{XY} and so on, we assume that

$$\operatorname{tr} \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1} = \operatorname{tr} \Sigma_{JI}\Sigma_{II}^{-1}\Sigma_{IJ}\Sigma_{JJ}^{-1}. \tag{4.1}$$

Further we suppose the existence of the true model (I_t, J_t) which consists of the smallest number of parameters satisfying (4.1) when $I_t = \{1, \ldots, q_t\}$ and $J_t = \{1, \ldots, p_t\}$. Also, let $(\underset{\sim}{x}_i', \underset{\sim}{y}_i')$ be N independent observations of $\underset{\sim}{z}'$ and put

$$S = \begin{pmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{pmatrix} = \sum_{i=1}^{N} \begin{pmatrix} \underset{\sim}{x}_i - \bar{\underset{\sim}{x}} \\ \underset{\sim}{y}_i - \bar{\underset{\sim}{y}} \end{pmatrix} \begin{pmatrix} \underset{\sim}{x}_i - \bar{\underset{\sim}{x}} \\ \underset{\sim}{y}_i - \bar{\underset{\sim}{y}} \end{pmatrix}' : (p+q) \times (p+q).$$

Consider the model (I,J) where $I = \{1, \ldots, q_1\}$ and $J = \{1, \ldots, p_1\}$. Corresponding to I and J, we partition S into 16 submatrices (S_{ij}); $i, j = 1, \ldots, 4$ as $S_{XX} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$: $q \times q$, $S_{XY} = \begin{pmatrix} S_{13} & S_{14} \\ S_{23} & S_{24} \end{pmatrix}$: $q \times p$, $S_{YY} = \begin{pmatrix} S_{33} & S_{34} \\ S_{43} & S_{44} \end{pmatrix}$: $p \times p$, S_{11} : $q_1 \times q_1$, S_{13} : $q_1 \times p_1$, S_{33} : $p_1 \times p_1$ and $S_{ij} = S_{ji}'$. Then the likelihood ratio test statistic of the model (I,J) and the full model is given by Fujikoshi (1982) as

$$\Lambda(I_f,J_f;I,J) \equiv -2\log\lambda = N\log\{|S_{22.1}||S_{44.3}|/\begin{vmatrix} S_{22.13} & S_{22.13} \\ S_{42.13} & S_{44.13}\end{vmatrix}\}, \qquad (4.2)$$

where

$$S_{ij.13} = S_{ij.1} - S_{i3.1}S_{33.1}^{-1}S_{3j.1} = S_{ij.3} - S_{i1.3}S_{11.3}^{-1}S_{1j.3},$$

$$S_{ij.k} = S_{ij} - S_{ik}S_{kk}^{-1}S_{kj}.$$

If $I \supseteq I_t$ and $J \supseteq J_t$ or $q_1 \geq q_t$ and $p_1 \geq p_t$, then (4.1) is true which yields $(\Sigma_{41.3}, \Sigma_{42.3}) = 0$ and $(\Sigma_{23.1}, \Sigma_{24.1}) = 0$. Hence, by the law of iterated logarithm, using $\ell_N = (N^{-1}\log\log N)^{1/2}$,

$$N^{-1}S_{22.1} = \Sigma_{22.1} + O(\ell_N), \qquad N^{-1}S_{44.3} = \Sigma_{44.3} + O(\ell_N), \quad \text{a.s.},$$

$$N^{-1}\begin{pmatrix} S_{22.13} & S_{24.13} \\ S_{42.13} & S_{44.13}\end{pmatrix} = \begin{pmatrix} \Sigma_{22.13} & \Sigma_{24.13} \\ \Sigma_{42.13} & \Sigma_{44.13}\end{pmatrix} + O(\ell_N) = \begin{pmatrix} \Sigma_{22.1} & 0 \\ 0 & \Sigma_{44.3}\end{pmatrix} + O(\ell_N), \quad \text{a.s.},$$

and

$$\Lambda(I_f,J_f;I,J) = N\log\{|\Sigma_{22.1}||\Sigma_{44.3}|/\begin{vmatrix} \Sigma_{22.1} & 0 \\ 0 & \Sigma_{44.3}\end{vmatrix} + O(\ell_N^2)\}, \quad \text{a.s.},$$

$$= O(\log\log N), \quad \text{a.s.}, \quad \text{if } I \supseteq I_t \text{ and } J \supseteq J_t. \qquad (4.3)$$

If $q_1 < q_t$ or $p_1 < p_t$ (which implies $I \not\supseteq I_t$ or $J \not\supseteq J_t$), then $(\Sigma_{23.1}, \Sigma_{24.1}) \neq 0$ or $(\Sigma_{41.3}, \Sigma_{42.3}) \neq 0$. Hence, $|\Sigma_{22.1}||\Sigma_{44.3}| > |\Sigma_{22.13}||\Sigma_{44.13}|$. Therefore,

$$\Lambda(I_f,J_f;I,J) \geq N\log\frac{|\Sigma_{22.1}||\Sigma_{44.3}|}{|\Sigma_{22.13}||\Sigma_{44.13}|} + O(\log\log N), \quad \text{a.s.}$$

$$\to +\infty, \quad (N \to \infty), \quad \text{a.s.} \qquad (4.4)$$

This discussion is applicable in the general case of $I \not\supseteq I_f$ or $J \not\supseteq J_t$. In this case let $I_f^* = I \cup I_t$ and $J_j^* = J \cup J_t$. When we restrict the variables of $\underset{\sim}{x}$ and $\underset{\sim}{y}$ as $x_i (i \in I_f^*)$ and $y_j (j \in J_f^*)$, the true model remains (I_t, J_t).

Recalling the definition (4.2) and using (4.3) and (4.4), we get

$$\Lambda(I_f,J_f;I_f^*,J_f^*) = O(\log\log N), \quad \text{a.s.},$$

$$\lim_{N\to\infty} N^{-1}\Lambda(I_f^*,J_f^*;I,J) > 0, \quad \text{a.s.}$$

Hence,

$$\Lambda(I_f,J_f;I,J) = \Lambda(I_f,J_f;I_f^*,J_f^*) + \Lambda(I_f^*,J_f^*;I,J) \to -\infty, \quad \text{a.s.},$$

$$\text{if } I \not\supset I_t \text{ or } J \not\supset J_t. \tag{4.5}$$

To prove (4.3) and (4.5), we need only to assume the finiteness of the first two moments of $\underset{\sim}{x}$ and $\underset{\sim}{y}$.

Now define $(\hat{I}_N,\hat{J}_N)$ which minimizes

$$G_N(I,J) = \Lambda(I_f,J_f;I,J) - (pq - \#I\#J)c_N$$

and

$$\tilde{I}_N = \{i \in I_f \mid G_N(I_{-i},I_f) > 0\}, \qquad \tilde{J}_N = \{j \in J_f \mid G_N(J_f,J_{-j}) > 0\}$$

where $I_{-i} = I_f - \{i\}$ and $J_{-j} = J_f - \{j\}$. Combining (4.3) and (4.5), we obtain

THEOREM 4.1. Let $\{\underset{\sim}{z}_i = (\underset{\sim}{x}_i',\underset{\sim}{y}_i')' : i = 1, \ldots, N, \ldots\}$ be i.i.d. with mean vector $(\underset{\sim}{\mu}_x',\underset{\sim}{\mu}_y')'$ and variance covariance Σ. Then $(\hat{I}_N,\hat{J}_N)$ and $(\tilde{I}_N,\tilde{J}_N)$ are strongly consistent estimators of the true model (I_t,J_t).

APPENDIX

Proof of Lemma 2.1. We prove that the (k,ℓ)-th element of $\sum_{i=1}^{N}(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}_N)\underset{\sim}{e}_i'$ is $O(\sqrt{N \log\log N})$, a.s., $(1 \le k \le q,\ 1 \le \ell \le p)$. Hence, we do not lose generality by assuming $q = 1$ and $Ee_1^2 = 1$. We prove

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)e_i = O(\sqrt{n \log\log n}), \quad \text{a.s.} \qquad \text{(A.1)}$$

To prove (A.1), we need to show

$$\sum_{k=1}^{\infty} P[\bigcup_{2^{k-1}<n\le 2^k}\{\sum_{i=1}^{n}(x_i - \bar{x}_n)e_i > K\sqrt{n \log\log n}\}] < \infty$$

for some positive constant $K > 0$. If $2^{k-1} < n \le 2^k$,

$$|\bar{x}_n - \bar{x}_{2^k}| = |n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_{2^k}) \le \{n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_{2^k})^2\}^{1/2} < \sqrt{M}.$$

Hence by the law of iterated logarithm,

$$(\bar{x}_n - \bar{x}_{2^k})\sum_{i=1}^{n} e_i = O(\sqrt{n \log\log n}), \quad \text{a.s.}$$

Thus we shall prove

$$\sum_{k=1}^{\infty} P[E_k] < \infty \qquad \text{(A.2)}$$

where $E_k = \bigcup_{2^{k-1}<n\le 2^k}\{\sum_{i=1}^{n}(x_i - \bar{x}_{2^k})e_i > K2^{k/2}\sqrt{\log k}\}$.

Define

$$e_{ik}' = \begin{cases} e_i & \text{if } |e_i| \le 2^{k/2}, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$P(E_k) \leq P(E_k') + P[\bigcup_{i=1}^{2^k} (e_i \neq e_{ik}')]$$

where

$$E_k' = \bigcup_{2^{k-1}<n\leq 2^k} (\sum_{i=1}^{n} (x_i - \bar{x}_{2^k}) e_{ik}' \geq K 2^{k/2} \sqrt{\log k}\}.$$

So,

$$\sum_{k=1}^{\infty} P[\bigcup_{i=1}^{2^k} (e_i \neq e_{ik}')] = \sum_{k=1}^{\infty} 2^k P[e_1 \neq e_{1k}'] = \sum_{k=1}^{\infty} 2^k P[|e_1| \geq 2^{k/2}]$$

$$= \sum_{k=1}^{\infty} 2^k \sum_{\ell=k}^{\infty} P[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}] = \sum_{\ell=1}^{\infty} P[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}] \sum_{k=1}^{\ell} 2^k$$

$$\leq \sum_{\ell=1}^{\infty} 2^{\ell+1} P[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}] \leq 2 \sum_{\ell=1}^{\infty} E e_1^2 I_{[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}]}$$

$$\leq 2Ee_1^2 = 2,$$

$$|Ee_{1k}'| = |E(e_{1k}' - e_1)| = E|e_1| I_{[|e_1| \geq 2^{k/2}]} \leq 2^{-k/2} Ee_1^2 = 2^{-k/2},$$

$$\sum_{j=1}^{n} |x_i - \bar{x}_{2^k}| |Ee_{ik}'| \leq \{n \sum_{i=1}^{n} (x_i - \bar{x}_{2^k})^2\}^{1/2} 2^{-k/2} \leq 2^{k/2} \sqrt{2M}$$

for large n. If we let $e_{ik} = e_{ik}' - Ee_{ik}'$ and $T_n = \sum_{i=1}^{n} (x_i - \bar{x}_{2^k})$, we obtain

$$P(E_k') = P[\bigcup_{2^{k-1}<n\leq 2^k} \{T_n \geq K 2^{k/2} \sqrt{\log k} - \sum_{i=1}^{n} |x_i - \bar{x}_{2^k}| |Ee_{ik}'|\}]$$

$$\leq P[\bigcup_{2^{k-1}<n\leq 2^k} \{T_n \geq K 2^{k/2} \sqrt{\log k} - 2^{k/2} \sqrt{2M}\}]$$

$$\leq P[\bigcup_{2^{k-1}<n\leq 2^k} \{T_n \geq K' 2^{k/2} \sqrt{\log k}\}] \equiv P[F_k], \quad \text{say},$$

where we can take a new constant $K' > 0$ if $K > 0$ is sufficiently large.

Therefore,

$$P(E_k') \leqq P[\sum_{i=1}^{2^k}(x_i - \bar{x}_{2^k})|e'_{2^k}| \geqq K'2^{k/2}\sqrt{\log k}\,]$$

$$\leqq 2\{1 - \phi(K'\sqrt{\log k})\} + C_0R_k$$

where $R_k = \sum_{i=1}^{2^k}|x_i - x_{2^k}|^3E|e_{1k}|^3/\{2^{3k/2}(1+\sqrt{\log k})^3\}$, where $\phi(x)$ is the standard normal distribution function and C_0 is a constant independent of n. The last inequality is due to Bikelis (1966). If $K' > \sqrt{2}$, then we know that

$$\sum_{k=1}^{\infty}\{1 - \phi(K'\sqrt{\log k})\} < \infty.$$

If $\gamma = 3$,

$$\sum_{k=1}^{\infty}R_k \leq C_1\sum_{k=2}^{\infty}\{k \log k\}^{-1} < \infty.$$

If $2 \leq \gamma < 3$,

$$\sum_{k=1}^{\infty}R_k \leq \Gamma\sum_{k=1}^{\infty}2^{(3-\gamma)k/2}E|e_{1k}|^3$$

$$< C_2\sum_{k=1}^{\infty}2^{-(3-\gamma)k/2}(\sum_{\ell=1}^{k}E|e_1|^3I_{[2^{(\ell-1)/2} \leq |e_1| < 2^{\ell/2}]} + 1)$$

$$\leq C_3\sum_{\ell=1}^{\infty}2^{-(3-\gamma)k/2}(E|e_1|^3I_{[2^{(\ell-1)/2} \leq |e_1| < 2^{\ell/2}]} + 1)$$

$$\leq C_3\sum_{\ell=1}^{\infty}E|e_1|^{\gamma}I_{[2^{(\ell-1)/2} \leq |e_1| < 2^{\ell/2}]} + C_4 \leq C_3E|e_1|^{\gamma} + C_4 < \infty$$

because $E|e_1|^{\gamma} < \infty$, where $C_1, \ldots, C_4$ are positive constants. Thus we complete the proof of (A.2).

REFERENCES

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory (B.N. Petrov and F. Czáki, Eds.). *Akadémiai Kiado*, Budapest, 267-281.

[2] AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, 30, 9-14.

[3] BIKELIS, A. (1966). Estimates of the remainder term in the central limit theorem. *Litovskii matem.*, sb. 6.3, 323-346.

[4] FUJIKOSHI, Y. (1982). A test for additional information in canonical correlation analysis. *Ann. Inst. Statist. Math.*, 34, 523-530.

[5] FUJIKOSHI, Y. (1983). A criterion for variable selection in multiple discriminant analysis. *Hiroshima Math. J.*, 13, 203-214.

[6] FUJIKOSHI, Y. and NISHII, R. (1986). Selection of variables in a multivariate inverse regression problem. *Hiroshima Math. J.*, 16, 269-277.

[7] HANNAN, E.J. and QUINN, B.G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc.*, *B*, 41, 190-195.

[8] McKAY, R.J. (1977). Simultaneous procedures for variable selection in multiple discriminant analysis, *Biometrika*, 64, 283-290.

[9] NISHII, R. (1986). Criteria for selection of response variables and the asymptotic properties in a multivariate calibration, *Ann. Inst. Statist. Math.*, 38, 319-329.

[10] RAO, C.R. (1973). *Linear Statistical Inference and its Applications.* John Wiley, New York.

[11] RISSANEN, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465-471.

[12] SCHWARTZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6, 461-464.

[13] ZHAO, L.C., KRISHNAIAH, P.R. and BAI, Z.D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.*, 20, 1-25.

END
JAN.
1988
DTIC